

AL-TP-1993-0003

AD-A262 865



**WHAT CHANGES OCCUR DURING
COMPLEX SKILL ACQUISITION?**

Kathy A. Hanisch

Department of Psychology
Iowa State University
W212 Lacomarcino
Ames, IA 50011

Charles L. Hullin

Department of Psychology
University of Illinois at Urbana-Champaign
603 East Daniel Street
Champaign, IL 61820

**HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352**

February 1993

Final Technical Paper for Period June 1987 - September 1992

Approved for public release; distribution is unlimited.

93-07325



3188

88

4 07 0-6

**AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS**

ARMSTRONG

LABORATORY

NOTICES

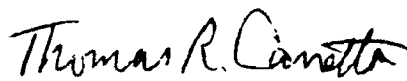
This technical paper is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

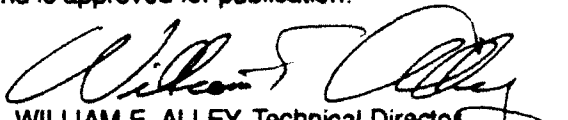
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this paper, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This paper has been reviewed and is approved for publication.



THOMAS R. CARRETTA
Contract Monitor


WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Research Division

ROGER W. ALFORD, Colonel, USAF
Chief, Manpower and Personnel Research Division

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 2104-0188	
<p>1. AGENCY USE ONLY (Leave blank)</p> <p>2. REPORT DATE February 1993</p> <p>3. REPORT TYPE AND DATES COVERED Final June 1987 - September 1992</p>				
<p>4. TITLE AND SUBTITLE What Changes Occur During Complex Skill Acquisition?</p>			<p>5. FUNDING NUMBERS C - F33615-87-C-0014 PE - 62205F PR - 7719 TA - 18 WU - 55</p>	
<p>6. AUTHOR(S) Kathy A. Hanisch Charles L. Hulin</p>				
<p>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Department of Psychology Iowa State University W212 Lacomarcino Ames, IA 50011</p> <p>Department of Psychology University of Illinois at Urbana-Champaign 603 East Daniel Street Champaign, IL 61820</p>			<p>8. PERFORMING ORGANIZATION REPORT NUMBER</p>	
<p>9. SPONSORING MONITORING AGENCY NAME(S) AND ADDRESS(ES) Armstrong Laboratory Human Resources Directorate Manpower and Personnel Research Division 7909 Lindbergh Drive Brooks Air Force Base, TX 78235-5352</p>			<p>10. SPONSORING MONITORING AGENCY REPORT NUMBER AL-TP-1993-0003</p>	
<p>11. SUPPLEMENTARY NOTES Armstrong Laboratory Project Scientist: Thomas R. Carretta, (210) 536-3942</p>				
<p>12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.</p>			<p>12b. DISTRIBUTION CODE</p>	
<p>13. ABSTRACT (Maximum 200 words)</p> <p>This paper presents and discusses the results of two separate but parallel studies of validity decrements of ability tests predicting skill acquisition and skilled performance. Hypotheses based on two explanations of the documented predictive validity decrements are tested. One explanation emphasizes transfer of training effects from learning and performing a complex task to the ability tests comprising similar elements of skills and knowledge. The second explanation emphasizes the effects of regression to the mean of ability measures in high ability groups that are selected, learned, and performed complex tasks. None of the hypotheses derived from these explanations for predictive validity decrements was supported. The results based on written pretest measures of ability replicated the basic phenomenon of decreasing predictive validities described in the literature. Analyses of predictive validities of computerized tests did not replicate the validity decrements across time and blocks of trials. Several analyses documented the advantages of exploiting assessments of performance taken during training as additions to prediction equations. The results of these analyses consistently showed that measures of performance obtained during training accounted for significant increments in variance beyond that accounted for by either pre- or post-training ability measures. The importance of this set of findings for practical, operational, solutions to the problems caused by predictive validity decrements is stressed.</p>				
<p>14. SUBJECT TERMS Complex skill acquisition Validity decay</p>			<p>15. NUMBER OF PAGES 34</p> <p>16. PRICE CODE</p>	
<p>17. SECURITY CLASSIFICATION OF REPORT Unclassified</p>	<p>18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified</p>	<p>19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified</p>	<p>20. LIMITATION OF ABSTRACT UL</p>	

CONTENTS

	<u>Page</u>
SUMMARY	1
INTRODUCTION.....	1
METHOD.....	5
Air Intercept Study	5
Subjects.....	6
Procedure.....	6
Measures.....	7
Air Intercept Training Performance Criteria	8
Flight Training Study.....	10
Subjects	10
Measures.....	11
Procedure.....	11
Flight Training Performance Criteria.....	12
Analyses.....	12
RESULTS.....	12
Air Intercept Study	12
Validity Decrement	14
Ability, Training, and the Air Intercept Task	15
Flight Training Study.....	19
DISCUSSION	21
REFERENCES	25

PREFACE

This research was accomplished under Contract F33615-87-C-0014 with Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division, Brooks Air Force Base. The authors thank Lloyd G. Humphreys for his guidance and suggestions during the early stages of the research design and implementation. Walter Schneider was helpful in providing the air controller's task that was used in the laboratory phase of the research. Fred Switzer, Sharon Furiya, Julie Olson, and Robert Kaiser provided assistance during the data collection stages of the research. Sherman Tsien provided much statistical analysis assistance. Special thanks are owed to the students, instructors, and administrators in the Aviation Institute for their willingness to cooperate in implementing numerous interruptions in their training procedures and for providing us with performance assessments.

Accession For	
NTIS	<input checked="" type="checkbox"/>
DND	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

WHAT CHANGES OCCUR DURING COMPLEX SKILL ACQUISITION?

SUMMARY

This report presents and discusses the results of two separate but parallel studies of validity decrements of ability tests predicting skill acquisition and skilled performance. Hypotheses based on two explanations for the documented predictive validity decrements are tested. One explanation emphasizes transfer of training effects from learning and performing a complex task to the ability tests comprising similar elements of skills and knowledge. The second explanation emphasizes the effects of regression to the mean of ability measures in high ability groups that are selected, learn, and perform complex tasks. None of the hypotheses derived from these explanations for predictive validity decrements was supported.

The results based on written pretest measures of ability replicated the basic phenomenon of decreasing predictive validities described in the literature. Analyses of predictive validities of computerized tests did not replicate the validity decrements across time and blocks of trials.

Several analyses documented the advantages of exploiting assessments of performance taken during training as additions to prediction equations. The results of these analyses consistently showed that measures of performance obtained during training accounted for significant increments in variance accounted for at all stages of training beyond that accounted for by either pre- or post-training ability measures. The importance of this set of findings for practical, operational, solutions to the problems caused by predictive validity decrements is stressed.

INTRODUCTION

Hulin, Henry, and Noon (1990) presented an extensive meta-analysis of the accumulated empirical, theoretical, and speculative literature in the area of predictive relations between ability measures and complex skill acquisition and skilled performance. They concluded that significant decrements in predictive validities could be expected across time or practice on the criterion tasks. These temporal validity decrements when ability measures were used to predict skilled performance across time were observed in educational, organizational, and experimental laboratory settings. The measures used to predict performance included, for example, general intelligence, job samples, and narrow measures of hand-eye coordination. Performance measures included, for example, the scientific productivity of engineers and scientists, ten-year performance of baseball players in the major leagues, and performance on a pursuit rotor task in a one-hour laboratory experiment. Observed validity decrements appeared to be general across measures, settings, and types of performance. Correcting observed empirical validity estimates for common statistical artifacts (e.g., changes in variance, reliability) resulted in an increase in the amount of validity decrement across time.

Hulin et al. (1990) also analyzed data using initial performance levels, as might be observed early in training, as predictors of performance in later time periods of training and post-training, operational performance as criteria. The conclusions reached on the basis of these stability, as opposed to predictive validity, coefficients, were that such measures displayed similar decrements to predictive validity coefficients across time. A corollary of these conclusions about stability coefficients is that the closer in time and practice to the late stages of operational performance criteria the predictors were assessed, the greater the expected predictive validity of the performance measures considered as predictors.

These latter conclusions appear particularly germane for organizations in which long periods of expensive practice and training are required before individuals have acquired a sufficient degree of skill to perform an organizational task. The results of the meta-analysis reported by Hulin et al. (1990) suggest that measures of performance taken during training, or even early in operational performance on a task, are significant predictors of later, operational performance levels and should be incorporated, along with ability measures, into any operational prediction equations. However, even these predictive relationships are not immune to the observed validity decrements and the later in training or performance the measures used as predictors can be taken, the better. To the extent that predictions depend substantially on test validities, expected validity decrements will adversely affect the predictions of performance late in training more than they will predictions of early operational performance.

The meta-analysis (Hulin et al., 1990) was mute with respect to the relative sizes of predictive validities based on traditional ability measures compared to assessments of performance during training or early in performance. However, these two general classes of predictors, ability measures and job samples, appear to be influenced by the same general psychological principals. This suggests that the distinctions between skills and abilities, when used as predictor measures, are more a matter of definitional convenience to researchers and theoreticians than a matter of any underlying fundamental differences. The similarities reinforce the conceptualization of human ability as an acquired repertoire of skills and knowledge possessed at a specific time by an individual (Adams, 1957; Alvares & Hulin, 1972, 1973; Hulin et al., 1990; Humphreys, 1960, 1973). The distinctions between job samples and most ability measures, then, reflect the specificity or generality of the definition and assessment of the ability/skill constructs that are assessed with skills falling at the specific end of the continuum and abilities falling at the general end of the continuum. Thus, an explanation for the observed validity decrements in ability measures may also explain the observed validity decrements of early or initial performance levels as well as the superdiagonal form of the resulting correlation matrix when independent performance assessments taken in sequential time periods are correlated. This superdiagonal matrix is a consequence of higher predictive validities for initial performance when predicting early performance and lower validities for predicting later performance. It is also possible that the processes underlying the observed validity decrements and the instability of performance measures across time are independent and unrelated; perhaps reflecting the result of several contributions to

change in rank orders of individuals across time as a function of time and practice on everyday tasks related to the criterion tasks being studied. It is unclear if the passage of time alone is sufficient to generate apparent instability in rank orders or if the time must be filled with practice on tasks related to the ability or performance dimensions being assessed.

The meta-analytic results raise several theoretical and conceptual issues that will be explored and debated over the next several years. The practical and applied implications, however, are more direct and immediate. What measures or combinations of measures will be most valid for predictions of performance on complex tasks after individuals have completed training and are performing their jobs or tasks in an operational environment?

This report summarizes attempts to test several theoretical explanations for the observed validity decrements. These theoretical explanations, if valid, could be used to design selection, placement, and training systems that exploit the robustness of ability measures as well as the practical, specific advantages of performance measures obtained during the early phases of training.

One theoretical explanation for the predictive validity decrement is based on the assumptions that abilities are neither capacities nor are they fixed by biological or early environmental factors. There is no evidence supporting assumptions about abilities as fixed capacities. All of the available evidence on the stability of ability measures suggests that changes in absolute amounts and rank orders of individuals in ability as a function of time or intervening practice on related tasks is the only constant in the equation. In this report, the focus shall be on changes in rank orders of individuals because this is the only change that influences correlations. Human abilities, no different than human physical measures, show changes that are revealed by decreasing correlations between ability scores separated by time or practice. Humphreys and his colleagues (e.g., Humphreys, 1960, 1973; Humphreys & Taber, 1973; Humphreys & Lin, 1977) have shown that independently observed measures of human ability obtained in sequential time periods are intercorrelated; the resulting correlation matrix displays a characteristic superdiagonal form. The adjacent correlations are large but become progressively smaller as the time between the observations becomes progressively longer. The important element of these data was that the correlations between ability measurements in the different time periods were reflecting changes in rank orders of individuals in terms of the abilities assessed. Rank order changes, not changes in group means across time, are the basis for the assumed changes in predictive validity.

The psychological mechanisms that control this process of ability change have not been well specified. Even though the distal cause of ability changes may be external, environmental events, the proximal cause must be a cognitive or psychophysiological change. The most likely explanation for the underlying changes in human ability would be the familiar process of transfer of training from tasks practiced and performed during the time between the ability assessments. To the extent that the tasks that were performed during the intervening time periods were composed of the

same or closely related tasks as those that compose ability measures, the practice on the related tasks should have significant transfer effects on the assessed abilities. If transfer of training were a valid explanation for the observed changes in correlations among ability measures, performance assessments, and predictive validity coefficients, then several experimental procedures become possible research tools. The amount and kind of related training could be manipulated. This should cause changes in ability in individuals who receive practice on related tasks relative to individuals that do not receive practice on related tasks.

A related explanation, based on individual change, is that normal daily experiences, maturation, development, aging, and regression to the mean account for the lack of stability in ability measures and predictive validity coefficients. The mechanism for change would not be transfer of training from interpolated practiced tasks to ability. It would be normal change associated with living from one time period to the next; little that we could do in the way of manipulated changes would influence the amount or speed of the changes in individuals. The impact of these naturally occurring changes and manipulations on individuals would be difficult to predict or evaluate without intensive and impractical ideographic studies of small samples of selected individuals. Further, institutionalizing or developing organizational interventions based on the observations may be impossible unless a small number of communalities among the normal, everyday influences are isolated and described.

There are obvious limitations to this latter explanation. The success of such programs as Project Head Start (Ramey & Ramey, 1990; Zigler, 1987) suggest that significant and massive changes, applied systematically and very early in life, can have an effect on individuals' abilities and performance later in life. The critical period for these interventions, however, may be very brief and the extent of the manipulations or changes may be so great that we cannot expect to duplicate the effects in normal, ethical, laboratory or field manipulations. Dunham's (1974) experimental findings notwithstanding, eight to twelve hours of practice on a supposedly novel task may have minimal impact on individuals; even the "artificial" laboratory tasks we develop may have a sufficient number of elements in common with normal work and play tasks that their impact on individuals in the experimental conditions is of marginal significance. Even learning to fly an airplane today may have less impact on individuals than it did as little as 20 years ago because of the extreme degree of control exerted by air traffic controllers (ATC's). This control leaves relatively little variation in flight path and approaches to unfamiliar airports up to the spatial orientation and spatial visualization skills of the student pilot. The effects of learning to fly on spatial ability that were observed by Alvares and Hulin (1972; 1973) were both statistically and practically significant. They also may have reflected a particular naturally occurring manipulation -- 1970's flight training -- that is no longer likely in a mid-1990's Airport Radar Surveillance Area and Terminal Control Area saturated environment. Thus, even this seemingly significant manipulation may not have the impact on basic spatial skills and abilities it once did because of the extensive vectoring and controlling by ATC. The explanation for the decrease in predictive validity coefficients, based on changes in rank orders of individuals in terms of their basic skills and abilities that were influenced by a transfer of training process, may be valid but difficult to test except in extreme

situations. Further, effective tests of the explanation may be limited to critical, developmental periods as suggested by Project Head Start.

Within these limitations, and the limitations imposed by the imprecision of the language of the transfer-of-training explanation for the predictive validity decrement, this report summarizes the results of both a field and laboratory experiment to evaluate the validity of the transfer of training and the regression to the mean explanations for the observed validity decrements. In the field experiment, the "manipulation" consisted of the individuals in the experimental group learning to fly an airplane during the first of a two-semester sequence leading to the private pilot's certificate. This course consists of a 15-week ground school; 6 hours in a Link GAT-I, a ground-based trainer; and approximately 23 hours of cockpit instruction in a Beechcraft C-19 or C-23 single-engine training aircraft. The students normally solo in the local traffic pattern or local practice area after approximately 11 or 12 hours of flight instruction but do no solo cross-country flights.

The laboratory experiment phase of this evaluation consisted of approximately 6 hours spent learning a complex task resembling that of an air traffic controller. A variation of this task was described by Schneider, Vidulich, and Yeh (1982) and Vidulich, Yeh, and Schneider (1983). Subjects were required to track multiple targets and vector interceptors into position to fire on the targets. Targets of different speeds at different locations and with different turning radii were presented throughout the task. In this study, two to three hours were spent on part-task learning and approximately three hours were spent on the complete task. Both experimental tasks are described more completely in the Method section below.

METHOD

Two separate studies were completed to examine the validity of two explanations for the decrements in predictive validity coefficients documented by Hulin et al. (1990). One based on transfer of training from related task performance and the other was based on regression to the mean of the population of highly selected subjects in organizations. The studies were carried on simultaneously. The Air Intercept (AIC) study was conducted over six semesters using college undergraduates learning a complex task that simulated an air traffic controller's job in the laboratory where ability levels could be controlled by means of random assignment of subjects to conditions.

A field study was also completed that used flight training students in their beginning course learning to fly an airplane. Separate descriptions of the procedures and measures used in each study are presented below.

Air Intercept Study

Individuals who participated in this study were selected based on their scores on a pretest screening battery. Many of the tests were chosen from a battery of tests described by French (1954) and are copyrighted by the Educational Testing Service,

Inc. (ETS). The pretest screening battery consisted of the following tests (described below): ETS Card Rotations, IPAT CAB-2-I, IPAT CAB-2-Mk, and Guilford-Zimmerman Spatial Orientation. The ETS Surface Development test was also administered but was not used as a screening device because of the time involved in scoring the test. The turnaround time following the administration of the screening tests for selecting subjects ranged from 24-48 hours.

Subjects

A sample of 97 experimental and 96 control subjects who completed all pretests, posttests, and the AIC task was obtained. Within the experimental group, 52 subjects were classified in the high ability group and 45 were classified in the random ability group on the basis of their pretest scores. Subjects were paid for their participation in the study with amounts varying by number of hours of participation. The standard rate was \$4.00 per hour with the possibility of those in the experimental group earning more money based on their performance on the AIC task. All of those performing in the top half on all four sessions of the AIC task were awarded bonuses as follows based on their overall performance levels on the AIC task: Best performance,=\$50, 2nd=\$25, 3rd=\$12.50, and the remaining performers in the top half of the distribution received an additional \$5. This bonus payment schedule was known ahead of time by the subjects and was intended to provide motivation and some degree of competitiveness similar to that found in normal organizations for task performance.

Procedure

A typical procedure in a semester involved screening approximately 155 subjects to select approximately 70. From their scores on the pretests, a stratified random sampling procedure was used to select a group that scored high on the tests and to randomly select a group that had a distribution of scores, including some high scores, approximating that found in the population. The tests were scored and the mean summed standardized scores were used to place individuals into one of three groups (1) experimental group - high ability, (2) experimental group - random ability, and (3) control group. Using the mean summed standard scores and the z-score, the distribution was divided as follows: $z\text{-score} < -.517$, 30%; $-.517 < z\text{-score} < -.006$, 20%; $-.006 < z\text{-score} < .612$, 20%; and $z\text{-score} > .612$, 30%. The high ability group was selected first from individuals with z-scores greater than .612. Individuals with high scores were randomly assigned using a random number table to one of the three groups based on the total number in the top group (this varied by semester and year of the study). The remaining individuals were divided into the various groups based on their z-scores and random assignment to group.

Once the individuals were assigned to the various groups there were two separate procedures for the control and experimental groups. Those assigned to the control group were required to complete 1 1/2 hours of computer tests (described below) consisting of: Sternberg Short Term Memory, Sentence-Picture Comparison, Mental Paper Folding, Arrival Time, Extrapolation, and Intercept. After completing these tests, the subjects were not contacted until the end of the semester and at that time completed the same written (from the screening battery) and computer pretests.

Individuals in the experimental group also completed the same computer pretests and were then scheduled for six hours of training and performance on the air intercept (AIC) task that simulated an air traffic controller environment. After completing the six hours, the subjects then completed the 1 hour of written and 1 1/2 hours of computer posttests. There was approximately a seven-week time interval between the pretest and posttest administrations for both the experimental and control groups.

Measures

A total of 11 tests were administered to subjects to allow for exploratory analyses. Two computer tests and four written tests were chosen for further evaluation based on their relations among the tests as well as their assumed relation to performance on the AIC task. A brief description of each test is given below.

Computerized tests. Two computer tests were selected for evaluation with respect to the training and task measures. The tests were completed on either an IBM Model 30 or 50 computer with peripheral joysticks. The computers were equipped with VGA cards and were coupled to monochrome screens that displayed 64 shades of gray.

(1) Extrapolation. For each trial, a curve began at the left and extended toward the right of the screen. The subject indicated at what point along a vertical line on the right side of the screen that the curve would hit if it continued its path. The curves were either a horizontal line, a parabola, or a sine wave. In addition, the trials differed in terms of the distance between where the curve ended on the screen and the vertical line (i.e., the distance the subjects had to extrapolate to the line). The subject marked his/her estimate on the vertical line using a joystick-controlled arrow. The performance measure for the 100 trials was the number correct based on the distance from the extrapolated point from the true point. Extrapolations that were within four centimeters were scored as correct.

(2) Mental Paper Folding. For each trial, the subjects were presented with a two-dimensional representation of a cube that had been cut apart and laid flat. Each diagram had arrows pointing to two edges of different squares. The subject pressed one of two keys indicating whether or not the two marked edges would meet if the two-dimensional diagram was folded into a cube. There were 60 items of various difficulty levels; performance was scored by the number of items correct.

Written tests (paper and pencil). Four written tests were chosen for evaluation based on the relations among the tests and their relations to the training and task measures.

(1) Guilford-Zimmerman Aptitude Survey (Spatial Orientation). This test was designed to measure a subject's ability to perceive changes in direction and position. Each of the 30 items consisted of a pair of pictures that showed the shore and the prow of a boat from the perspective of someone in the boat. The subject

determined which of five schematic diagrams matched how the position of the boat had changed from the first picture to the second picture. Performance was scored by the total number of correct responses in a 5-minute period.

(2) IPAT CAB-Mk. This test measured knowledge about mechanical facts and principles. Each of the 18 items presented either a picture or a description that asked about specific mechanical knowledge. Performance was measured by the number correct in a 6 1/2 minute period.

(3) IPAT CAB-I. The ability to infer a rule from patterns of letters was measured by this test. Each of the 12 items consisted of five 4-letter strings. Four of the five strings followed a certain rule (e.g., alternating vowels and consonants with the consonants in alphabetical order). The subject was to mark the string that did not follow the rule. Performance was measured by the total number of correct responses in a 6-minute period.

(4) ETS Surface Development Test. This test was designed to measure a subject's ability to visualize how a piece of paper could be folded to form a 3-dimensional object. For each of 12 sets of items, a 3-dimensional object was presented as if it were cut apart and laid flat. Dotted lines indicated where the flattened object should be folded to form the 3-dimensional object that was pictured next to it. Each item consisted of a numbered side of a flattened object with five items per object. The subject was to figure out which of the lettered edges on the 3-dimensional object were the same as the numbered edges on the flattened object. Performance was measured by the total number of correct responses in a 12-minute period.

Air Intercept Training Performance Criteria

Air Intercept Training Tasks. The following describes the operation and execution of the AIC training task that experimental subjects completed in blocks of 90 minutes over approximately 6 hours on an IBM Model 30 or 50 computer. The model 30's had MCGA graphics cards. The maximum resolution on an MCGA card is 640 (horizontal) X 200 (vertical) pixels. The model 30's had monochrome monitors. The model 50 had color but it ran the software in the same resolution in black and white as the model 30's did. At the outset of the task, subjects were given a representation of a compass, a card with a circle on it divided into angle sections. They were to use this as an aid in identifying the heading of the plane that was presented to them on the computer screen. All heading references were made in terms of the angle direction as opposed to standard ordinal directions (e.g., 90 degrees rather than east). The basic training task operations, as the subjects were presented them by the computer, are given below. In addition, they were given a brief introduction to the computer commands and keys that they needed to use throughout the task. There were a total of seven training tasks. Within each task, subjects were given five practice trials that were not scored and were given feedback on their performance within each task. Each of these seven training tasks contained trials that emphasized elements of the final AIC task. The tasks were also arranged roughly in order of increasing complexity and similarity to the final task.

(1) Identify Heading of Flightpath. Subjects were to visualize and then identify the heading of an aircraft presented to them on the computer screen. They were instructed to use the keyboard to enter the heading and were given feedback as to the accuracy of their value. Measures included heading error and reaction time across 30 trials.

(2) Reciprocal Heading of Aircraft Path. Subjects were to calculate the reciprocal heading of the aircraft path. In addition, they were to enter the turn direction (Left [L] or Right [R]) as specified on the screen from the perspective of the plane. This left/right turn direction was to acquaint the subjects with the turns from the perspective of the pilot and to become familiar with the appropriate keys. Measures of performance included heading error, direction error, and reaction time across 30 trials.

(3) Bearing and Range of the Radar Blip from the Fighter. Two objects were on the screen -- a moving symbol fighter plane and a symbol for a radar blip. The subjects' task was to estimate the bearing and the range of the radar blip from the fighter. Bearing was the heading the subject (the fighter) would take to get to the radar blip; the range was the distance of the radar blip from the fighter. The subjects' answers were to be within 10 degrees (bearing) and 1 mile (range). Subjects entered the bearing and range and were given feedback as to their accuracy. Measures of performance included heading error, angle error, and reaction time across 30 trials.

(4) Hit a Stationary Point at a Specified Heading. The subjects were presented a heading and were to visualize the fighter turning towards that heading so that it would intercept the radar blip. Their task was to estimate the point, equivalent to both the time to turn and the point in space to begin the turn, where the fighter would need to turn to hit the blip given the fighter would turn to the specified heading. When the fighter reached the point that it should turn, subjects pressed a key. A square would appear on the screen after they had pressed the key to mark the correct turn point. The fighter would turn when and where they had indicated and the distance from the correct point and their turn would be shown. The measure of performance was distance error, distance of correct turning point from the point selected by the subject, across 30 trials.

(5) Visual Identification of the Intercept Point. A fighter, an unidentified contact labeled a "bogey," and cursor were on the screen in this task. The fighter and the bogey moved at the same speed and the bogey did not change direction. The subjects were to visualize a line from the bogey in the direction of its heading. A target crossing angle (TCA) was presented and told them at what angle the fighter should intercept the bogey (a sheet was given to subjects that defined and described what the TCA was). They were then to use the cursor (by moving the arrow keys) that would move only on the path of the bogey. They were to place the cursor at the point at which the fighter would intercept the bogey at the given angle (the fighter will turn to hit the bogey and the bogey will continue on its path) and then press a key. A square would appear where the cursor was and a radar blip would appear where the subject should have placed the cursor (the real point of intercept). The measure of performance was distance error, the difference between the real point of crossing and that selected by the subject, summed across 30 trials.

(6) Intercept Calculation. This task involved the fighter and bogey on the screen. The subjects' task was to visualize the heading of the bogey (it would not change) and the given TCA on the heading of the bogey. They were then to type in the direction and the heading the fighter must take to hit the bogey with the given TCA. Feedback as to how close they came to hitting the bogey was presented on the screen. Measures included heading error, direction error, distance error, and reaction time across 30 trials.

(7) Advanced Bogey Intercept. This task was very similar to Intercept Calculation except that subjects were no longer given the target crossing angle, but had to visualize it. The vapor trails from the aircraft were no longer given (they had been given in all earlier tasks where relevant) so the subjects had to visualize the direction of the bogey and fighter from their movement alone. Subjects visualized the heading of the bogey, typed in the direction and the heading the fighter must take to hit the bogey. Feedback was given on the screen to show how close the subject came to hitting the bogey. Measures included heading error, direction error, distance error, and reaction time summed across 87 trials.

Air Intercept Task -Advanced Bogey Intercept with Stranger Alert. The final task was similar to the last training task, advanced bogey intercept. There was one additional feature on this task. Subjects were to estimate when the fighter should turn and the heading that the fighter should turn to so that it intercepted the bogey. In addition, the subjects needed to pay attention to other objects on the screen in addition to the bogey and the fighter planes. There were several radar blips that were called "strangers." In addition to entering the direction and heading of the fighter, the subject needed to keep track of the strangers on the screen. Whenever a stranger came within 5 miles of the fighter plane, the subject needed to press a key once. The subject was instructed to press a key once for each additional stranger that came within 5 miles of the fighter plane. Measures included heading error, direction error, distance error, and reaction time summed across 240 trials.

The final AIC task was scored in blocks of 20 trials for a total of 12 measures of performance. This was done to allow an examination of relative stages (i.e., early, middle, late) of performance on the operational task. All of the measures used in the analyses for the training and final task consisted of summing the various measures assessed for each segment of performance across the trials. The practice trials on the seven partial tasks and the final AIC task took subjects between five and six hours in blocks of 90 minutes over a two-week period to complete.

Flight Training Study

Subjects

Subjects in the experimental group were 98 undergraduate students in an introductory first semester aviation class. This was the first semester of a two-semester sequence of courses leading to the private pilot's certificate. The control group was comprised of 142 first year business and engineering students. Control group

subjects were solicited from the Business and Engineering schools to attempt to control for the mean American College Testing program scores and, to a lesser degree, interests of the two groups. All subjects were paid a nominal fee for participation in the ability testing portion of this study. Subjects who completed all stages of the study were entered into random drawings for additional monetary and prize awards.

Measures

Pre- and post flight-training ability was measured using a battery of four written and six computerized tests. The written tests were: Bennett Mechanical Comprehension, Guilford-Zimmerman Aptitude Survey (Spatial Orientation), ETS Choosing a Path, and ETS Maze Tracing. The computerized tests were the same as those used in the AIC study. Based on the relations among the tests and tasks, one computer and three written tests, were chosen for further evaluation: Extrapolation, Guilford-Zimmerman, Bennett, and Maze Tracing.

Written tests. All but the Guilford-Zimmerman written test used in the flight training study were unique to that study; the tests used in this study are described below.

(1) Bennett Mechanical Comprehension Test. This test was designed to measure the ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations. It consisted of 68 items that required the subject to look at a picture and answer questions about the physical and mechanical properties illustrated. The items are typically simple and do not require esoteric or highly specialized knowledge; they are based on a general understanding of mechanical principles that could be gained from normal, everyday, activities and observations. Performance was measured by the total number correct in a 15 minute period.

(2) Maze Tracing Test. This test was designed to measure an individual's ability to quickly find the correct path through a maze. For each of 48 mazes, the subject was asked to draw a pencil line through each maze without crossing any printed lines. Performance was measured by the total number of correct mazes solved in 6 minutes.

Procedure

The battery of ability measures was administered twice to all subjects, separated by 14 weeks. For subjects in the experimental group, initial flight performance was measured after approximately 10 hours of flight experience. Intermediate flight performance was measured after approximately 18 hours of experience. Both initial and intermediate flight performance were assessed in a flight simulator developed at the University of Illinois and approved for flight training and logging of flight time. Final flight performance was measured in an airplane after approximately 23 hours of flight experience. All flight performance measures were taken after administration of the ability pretest measures and before administration of the ability posttest measures.

Flight Training Performance Criteria

Assessments of initial and intermediate flight performance were based on performance in the flight simulator. Final flight performance was measured in actual check rides in a Beechcraft C19/23, single engine, 180 HP, planes equipped for flight under instrument flight rules. Differences between the C19 and C23 are negligible. Flight performance measures were taken by trained observers who were present with the student in the simulator or aircraft; the procedure duplicates closely the standard FAA check ride procedure. Observers recorded several instrument readings at specified times on a rating form developed for the study. This rating form attempted to reduce the mental workload of the observer by asking him or her to check altimeter, vertical speed indicator (VSI), bank, airspeed, ball position, heading, etc., and record the readings on a standard form at specified times during a maneuver. The check pilots and observers were relieved of the mental workload of integrating quality of performance across an entire maneuver or check ride; they only had to take and record observations of specified flight parameters at specified times during a maneuver. Based on the observers' recordings of instrument readings, two overall scores reflecting the quality of flight were obtained. The first score measured the accuracy of each of the individual maneuvers performed during the flights; the second score measured the accuracy of specific flight parameters (e.g., heading, airspeed, altitude). Because both the maneuvers score and the flight parameter scores were comprised of the same instrument readings accumulated and summed in different ways, the scores were not independent. All flights were done with a view restricting device being worn by the pilot and consisted of standard instrument flight maneuvers, climbs, descents, radial intercepts, holds, standard rate turns, climbing and descending turns, radar vectoring, and one instrument approach.

Analyses

Changes in mean ability scores in experimental groups, as compared to the control groups, were tested by means of repeated measures analyses of variance. Relations of ability scores to performance levels, changes in relations across trials, independent relations involving ability measures and training performance on the one hand and performance on the criterion tasks were examined by means of least squares and hierarchical regression analyses. Changes in variance accounted for in the criterion tasks entering the predictor variables in different predetermined orders were used to test the incremental validity of each set of measures. Several post hoc analyses were conducted to explore relations among the measures.

RESULTS

Parallel analyses were conducted wherever possible across the AIC and Flight Training studies. The results are presented separately by study.

AIC Study

All of the variables in this study were standardized to have a mean of zero and a standard deviation equal to one across the combined experimental and control groups

in the AIC study. This transformation preserves any pre- or posttest differences between the experimental and control groups. Further, all relevant information is preserved in the transformed measures because changes in the means, say in the experimental group, are tested relative to any changes in the control group across the pre- and posttest assessments. The means and standard deviations on the pretests and posttests for the experimental and control groups, as well as the random and high ability groups within the experimental group, are shown in Table 1. The experimental and control group means on the ability measures were not significantly different ($p > .05$); the means for all of the pretests and posttests for the experimental random and high ability groups were significantly different ($p < .05$). The expected difference in the mean ability level between the two experimental groups, high ability and random, was achieved.

Table 1. Pretest and Posttest Means and Standard Deviations for the AIC Study Groups

Tests	Experimental Group Mean/SD	Control Group Mean/SD	Experimental Random Ability Group Mean/SD	Experimental High Ability Group Mean/SD
Extrapolation-Pretest	.10/1.05	-.10/0.95	-.30/1.04	.44/0.93
Extrapolation-Posttest	.08/1.02	-.08/0.98	-.33/1.01	.46/0.87
Mental Paper Folding Pretest	-.02/1.02	.02/0.99	-.28/1.04	.20/0.95
Mental Paper Folding Posttest	.12/0.96	-.11/1.03	-.12/0.96	.32/0.92
Guilford-Zimmerman Pretest	.03/1.00	-.04/1.01	-.49/0.92	.49/0.83
Guilford-Zimmerman Posttest	.10/0.98	-.10/1.01	-.50/0.96	.63/0.64
IPAT CAB-MK Pretest	.08/1.07	-.08/0.91	-.49/1.15	.58/0.70
IPAT CAB-MK Posttest	.07/1.08	-.08/0.91	-.48/1.13	.54/0.77
IPAT CAB-I Pretest	.01/0.96	-.02/1.04	-.45/1.08	.41/0.63
IPAB CAB-I Posttest	.08/0.84	-.07/1.14	-.30/1.08	.40/0.32
Surface Development Pretest	.09/0.98	-.10/1.01	-.33/1.10	.46/0.69
Surface Development Posttest	.03/0.97	-.03/1.04	-.42/1.23	.42/0.36
Sample Size	97	96	45	52

Note: None of the experimental versus control group mean differences are statistically significant; all of the experimental random versus high ability group mean differences are statistically significant ($p < .05$).

The correlations, including stability coefficients from time 1 to time 2, among the pretest and posttest measures are shown in Table 2. The pretest/posttest stability coefficients are italicized, bold values and range from .59 (extrapolation) to .83 (IPAT CAB-Mk). All of the correlations among the pretests and posttests are statistically significant ($p < .05$) and, as expected, positively related.

Table 2. Correlations Among the Pretest and Posttest Ability Measures for the AIC Study

<u>Tests</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
Extrapolation Pretest	--											
Extrapolation Posttest	59	--										
Mental Paper Folding Pretest	31	28	--									
Mental Paper Folding Posttest	24	24	70	--								
Guilford-Zimmerman Pretest	32	28	36	26	--							
Guilford-Zimmerman Posttest	44	44	39	30	78	--						
IPAT CAB-MK Pretest	24	32	33	30	47	52	--					
IPAT CAB-MK Posttest	30	28	32	23	48	50	83	--				
IPAT CAB-I Pretest	31	36	31	26	23	38	31	27	--			
IPAT CAB-I Posttest	25	27	39	33	34	45	28	23	59	--		
Surface Development Pretest	44	45	51	34	48	57	53	54	52	43	--	
Surface Development Posttest	35	39	58	46	53	60	56	55	48	46	78	--

Notes: N = 191-196; decimals omitted from correlations; italicized, bold values are stability coefficients; all correlations are statistically significant ($p < .05$).

Validity Decrement

Figure 1 summarizes the trends in predictive and postdictive validities across the 12 blocks of trials on the final AIC task for the complete set of written and computerized tests. The regression of the predictive validities of written pretest ability measures for blocks of trials in the AIC task had a slope of $-.77$ ($p < .01$). The negative slope of the written pretests validities onto the blocks of trials replicates the documented decrement in predictive validities (Hulin, et al., 1990). The regression of postdictive validities of the written posttests for the same blocks of trials had a slope of $-.63$ ($p < .05$); this slope was not significantly different from the slope of $-.77$ for the written pretests although the power of this test to detect inferences is not great. The regression of the predictive validities of the computerized pretests onto the blocks of trials revealed no trend. The slope was $.07$ ($p > .05$) for the computerized pretests; it was $-.30$ ($p > .05$) for the

computerized posttests. These results indicate that the written pretests and posttests displayed the expected validity decrement throughout the blocks of trials on the AIC task. The computerized tests, even though selected from the same general domains as the written tests, did not display any trends in validities across the blocks of trials.

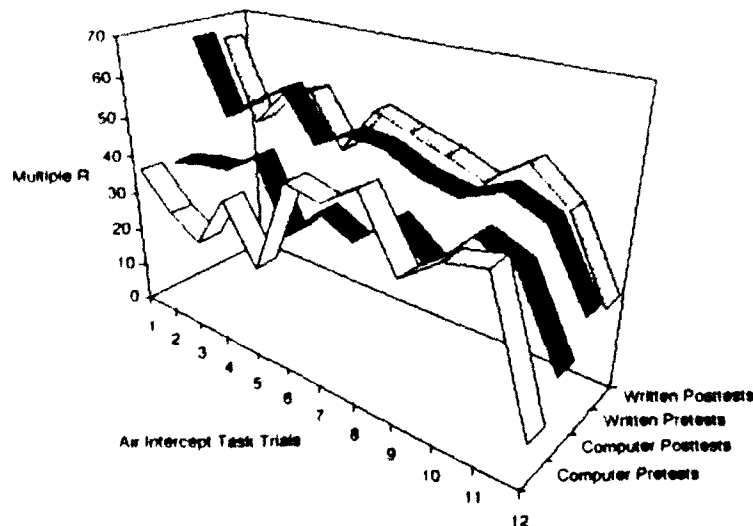


Figure 1. Regression of Written and Computer Tests on the AIC Task

Ability, Training, and the AIC Task

To examine what measures were valid for predictions of performance on the final AIC task after subjects had completed the seven training tasks, and to test the transfer of training explanation for validity decrements, hierarchical multiple regressions on the 12 blocks of trials for the final AIC tasks were completed using the pretests, posttests, and measures of performance taken during training that were described above. Four predictions and one overall measure of performance for each block of trials were analyzed. Tables 3, 4, and 5 focus on the change in R in predicting performance over time on the final AIC task as a function of the pretests, training tasks, and posttests.

Table 3 presents the changes in multiple correlations in predicting AIC task performance across 12 blocks of trials as a function of entering either the pretests and then posttests, or the reverse, into the multiple regression equation. The results are presented separately by ability group. The change in R when posttests were added to the equation after pretests had been entered is statistically significant in only the first three trials for the random ability group. There are two trials (6 and 8) for the high ability group where the change in R was significant when pretests were added to the equation after posttests had been entered.

These results indicate that the posttest measures, taken following training on the criterion task, do not add significantly to the variance accounted for after the pretest measures were entered. If an explanation of the validity decrement based on transfer

of training were valid, there should be a stronger relationship between posttest ability measures and final performance than between pretest ability measures and final performance measures. The reverse, stronger relations between pretest ability measures and initial performance than between pretest measures and final performance measures, should also be true. The rank orders of the subjects on final performance measures and posttest ability measures should be more similar because the changes induced in individuals' abilities by criterion task practice would be reflected in both posttest measures and final task performance. Similarly, pretest measures and initial criterion task performance should reflect the rank orders of individuals on initial abilities, the abilities that influence initial performance levels. The results, shown in Table 3, do not support this explanation.

Table 3. Multiple Correlations, R's, for Hierarchical Regressions of Pretests and Posttests Across the Air Intercept Task Trials

<u>Trials</u>	<u>Step 1 Pretests</u>		<u>Step 2 Posttests</u>		<u>Step 1 Posttests</u>		<u>Step 2 Pretests</u>	
	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>
1	30	59	31	65	15	64	31	65
2	27	48	27	57	20	57	27	57
3	24	48	27	65	26	62	27	65
4	28	54	28	58	24	58	28	58
5	22	37	23	39	14	39	23	39
6	48	43	48	46	29	46	48	46
7	10	52	24	52	08	47	24	52
8	40	52	43	53	18	51	43	53
9	20	54	22	55	20	52	22	55
10	20	57	24	57	24	55	24	57
11	35	40	37	40	34	39	37	40
12	13	29	14	40	13	38	14	40

Notes: Bold, italicized values indicate a significant ($p < .05$) increase in Multiple R from Step 1 to Step 2; decimals omitted from correlations.

Table 4. Multiple Correlations, R's, for Hierarchical Regressions of Pretests and Posttests Across the Air Intercept Task Trials

<u>Trial</u> s	Step 1 Pretests		Step 2 Posttests		Step 1 Posttests		Step 2 Pretests	
	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>
1	36	64	38	67	30	59	38	67
2	59	53	59	54	27	48	59	54
3	43	64	43	64	24	48	43	64
4	52	58	52	61	28	54	52	61
5	39	54	39	54	20	37	39	54
6	47	48	54	49	48	43	54	49
7	41	57	44	60	10	52	44	60
8	54	58	55	60	40	52	55	60
9	46	65	46	66	20	54	46	66
10	51	58	53	62	20	57	53	62
11	46	43	47	45	35	40	47	45
12	34	61	35	62	13	29	35	62

Notes: Bold, italicized values indicate a significant ($p < .05$) increase in Multiple R from Step 1 to Step 2; decimals omitted from correlations.

Table 5. Multiple Correlations, R's, for Hierarchical Regressions of Training and Posttests Across the Air Intercept Task Trials

<u>Trial</u> s	Step 1 Pretests		Step 2 Posttests		Step 1 Posttests		Step 2 Pretests	
	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>	<u>High Ability</u>	<u>Random Ability</u>
1	36	64	37	68	13	64	37	68
2	59	53	60	58	20	57	60	58
3	43	64	43	67	25	62	43	67
4	52	58	52	61	24	58	52	61
5	43	54	44	55	14	39	44	55
6	47	48	47	49	28	46	47	49
7	40	57	47	57	07	47	47	57
8	54	58	55	59	15	51	55	59
9	45	65	46	65	21	52	46	65
10	50	58	50	60	24	55	50	60
11	45	43	47	44	33	39	47	44
12	34	61	34	62	14	38	34	62

Notes: Bold, italicized values indicate a significant ($p < .05$) increase in Multiple R from Step 1 to Step 2; decimals omitted from correlations.

To test this same explanation from a different perspective, six repeated measures analyses of variance were done to test for the significance of any changes between the pretests and posttests. The pretest and posttest scores were treated as repeated measures and the control and experimental groups as the manipulation. In the analyses, involving the six tests, there was one significant ($F = 6.13, p < .05$) time-by-group interaction involving the Mental Paper Folding test. The experimental group significantly increased its mean from pretest to posttest as a function of the practice on the AIC task while the control group mean decreased. None of the other interactions involving the other five tests was significant. Given the number of tests involved, this singular time-by-group interaction will not be further interpreted. These results indicated that an explanation for the observed predictive validity decrements in the literature based on a transfer of training explanation cannot be supported on the basis of these data.

Six repeated measures analyses of variance involving the high versus random groups, indicated that all six of the tests were significantly different between the two groups. None of the time effects was significant and none of the time-by-group interactions was significant. These results indicate that an explanation for the documented predictive validity decrements based on a differential regression to the mean hypothesis was not supported.

The results in Table 4, where training performance assessments and the pretest measures were entered to examine the change in R when predicting task performance, indicate that adding training to the pretests scores results in significant increases in R in 18 out of 24 comparisons across the high and random ability groups. Adding the pretests to the equation when training has already been entered results in a significant increase in R in only 1 out of 24 comparisons.

Results similar to the pretest-training findings are obtained when training and the posttests measures are used to predict performance on the AIC trials. Table 5 presents these results.

None of the equations where posttests were added to training measures resulted in significant increases in R; 18 out of 24 equations predicting task performance when posttests were entered first and then training measures were added resulted in significant increases in R across the groups.

These results attest to the importance of considering performance during training as a predictor of final performance levels on these experimental tasks. Performance during training significantly increased the validity of the equation whether pretest or posttest ability measures were entered into the equation first. Entering performance measures obtained during training into the equation first and then ability measures, whether pretest or posttest, did not result in a significant increase in the variance accounted for in final task performance. The results were not restricted to either the random or the high ability groups. This generality across both groups strengthens the interpretation of the phenomenon as being highly relevant to organizational settings dealing with highly selected subjects. These results document the advantages of

considering all information, particularly training, related to final performance on a complex task in our final predictions. Although initial selection procedures that must be made before any training is undertaken and performance measures are obtained cannot benefit from a consideration of training measures, any initial predictions made on the basis of pre-training ability measures can be updated as soon as the first training is undertaken and performance is assessed.

Flight Training Study

Table 6 displays the results of mean changes in test scores observed in the flight study from the pretests to the posttests. There are several trends in the data revealed in this table. The first is that although the control group had higher means than the experimental group on three of the four tests, none of these differences was statistically significant. However, the control group had a significantly lower mean than the experimental group for the Bennett Test of Mechanical Comprehension. These results very likely represented a combination of the alpha level in operation and small differences among groups of students enrolled in the different curricula from which the subjects were sampled.

Table 6. Pretests and Posttests Means and Standard Deviations for the Flight Study Groups

<u>Tests</u>	<u>Experimental Group Mean/SD</u>	<u>Experimental Group Mean/SD</u>
Extrapolation-Pretest	-.05/0.94	.07/0.98
Extrapolation-Posttest	-.01/0.94	.03/1.04
Guilford-Zimmerman Pretest	-.04/1.05	.03/0.99
Guilford-Zimmerman Posttest	-.07/0.98	.07/1.01
Bennett Pretest	.24/1.10	-.15/0.94
Bennett Posttest	.11/1.06	-.06/0.96
Maze Tracing Pretest	-.10/1.03	.17/0.98
Maze Tracing Posttest	-.18/0.99	.15/0.98
Sample Size	98	142

Notes: The Bennett Pretest, Maze Tracing Pretest, and Maze Tracing Posttest means are significantly different across the experimental and control groups.

A second important finding was that none of the differences from pretests to posttests for the experimental groups was significant. This finding replicated the finding from the AIC experiment and further failed to support the transfer of training explanation for observed predictive validity decrements.

Table 7 presents the intercorrelations among the pretest and posttest ability measures administered to the flight students as well as the stability coefficients for the tests over the 14-week interval of the study. The stability coefficients and the intercorrelations among the tests were as expected.

Table 7. Correlations Among the Pretest and Posttest Ability Measures for the Flight Study

<u>Tests</u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>
Extrapolation Pretest	--										
Extrapolation Posttest	49	--									
Guilford-Zimmerman Pretest	21	20	--								
Guilford-Zimmerman Posttest	20	20	56	--							
Bennett Pretest	07	08	42	36	--						
Bennett Posttest	12	18	41	50	69	--					
Maze Tracing Pretest	17	08	37	28	27	29	--				
Maze Tracing Posttest	13	06	29	29	15	23	76	--			
Parameters-Simulator Ride 1	13	24	06	09	-02	-20	01	01	--		
Parameters-Simulator Ride 2	15	16	16	15	12	02	16	15	52	--	
Parameters-Airplane Checkride	-06	09	12	-02	19	02	04	02	52	64	--

Notes: N = 47-242; decimals omitted from correlations; italicized, bold values are stability coefficients; italicized values are statistically significant ($p < .05$).

The correlations between the ability measures and the flight parameter scores derived from the flight performance assessments were disappointing. There was little communality between the tests and flight performance during learning as assessed in this study. These results cannot be attributed to the lack of stability of the ability measures, the lack of stability in the flight performance measures, or the homogeneity of the scores in the two sets of measures. Other statistical artifacts, such as range restrictions, are not ruled out. These results may simply indicate a lack of significant relations among the ability measures and assessments of skilled performance.

DISCUSSION

This report summarizes the results of the tests of two theoretical explanations of the decrements in predictive validities that are observed when ability measures are used to predict performance on skilled, cognitive or psychomotor tasks. One explanation, offered 20 years ago by Alvares and Hulin (1972; 1973), is based on a transfer of training process. This explanation began with a definition of human ability that emphasizes that abilities consist of the current repertoire of skills and knowledge possessed by an individual at some point in time. The acquisition of these skills likely has a biological substratum, but is also heavily influenced by the environment and experiences of individuals at different, critical times in their lives. Skills and knowledge, and thus the estimates of abilities that are obtained by different, arbitrary combinations of skills and knowledge are not fixed; they are influenced by learning and experience and thus should be expected to change as a result of relevant training and experience. The distinction between skills and knowledge on the one hand and ability on the other hand is neither clear nor consistent. The distinction is one that is of greater convenience to the theoreticians and researchers than is required by the empirical data. The narrower and more specific (in the factor analytic sense of these terms) the assessment of skills or knowledge, the more likely the resulting variable or construct will be labeled a skill or specific knowledge. The broader and more general the assessment and combination of skills and knowledge, the more likely the resulting measure will be labeled as an ability. The most general ability would be general intelligence; the most specific skill might be a specific measure of hand-eye coordination or knowledge about tasks used in a particular skilled trade. This approach to cognitive and psychomotor human abilities, suggests that we should expect lawful change in the amounts of abilities and rank orders of individuals along the various ability continua that are created and studied by researchers.

Absolute amounts and relative standings of individuals along these continua are, at least partially, the result of both direct practice on the skills and knowledge comprised by the abilities and transfer effects from practice on tasks related to the abilities. These related tasks would likely contain elements in common with the skills and knowledge composing the relevant abilities; these common elements form the basis for assumed or demonstrated relations between abilities and skills/knowledge.

According to Humphreys' (1973) explanation for predictive validity decrements, relations between initial measures of ability and early performance on the task in question are the result of similar elements among the tasks and abilities. However, practice and performance on the task, as well as the necessary passage of time with all of its attendant but unspecified influences on individuals, alters the amounts of the skills and knowledge, and thus ability, possessed by the individuals in the selected sample. These changes in skills/knowledge and abilities do not decrease the relationship between ability and performance on the task. They do, however, reduce the relationship between the initial measures of ability and later assessments of performance. The later assessments of performance are still related to ability but the initial measures of ability are a relatively poor reflection of the skills/knowledge used to perform the task during the later stages of performance. Measures of ability taken

periodically during practice and training on the task or after extensive practice on the task would be better indications of the abilities used by individuals during the later stages of practice.

Based on the transfer of training explanation, we would expect to observe strong and significant relations between initial measures of ability and early performance on a task but lower relations between these same initial measures of ability and performance on a task after extensive practice and training on the task. These are, of course, the empirical results the explanation was intended to account for. We would also expect strong relations between measures of ability taken at the end of training or after extensive practice on a task and performance during the later stages of the task. The data in neither of these studies supported the transfer of training explanation.

A related test of the same explanation was based on changes in ability measures as a function of performance on the criterion task. Multivariate and univariate analyses of variance testing this hypothesis examined the significance of time-by-group interactions in two (experimental versus control group) by two (pre-training versus post-training assessments of abilities) repeated measures analyses. Only one of the comparisons in these analyses revealed a significant interaction in which the experimental group increased its ability scores more than the control group as a function of practice on the task. These analyses did not support the transfer of training explanation for the predictive validity decrements.

Additional related analyses addressed issues raised by the transfer of training explanation. For example, practice on a criterion task should decrease the stability/reliability coefficients of ability measures related to the criterion task more for the experimental groups than for the control groups who received no practice on the criterion task. The stability/reliability coefficients of the ability measures for the experimental group were not significantly smaller than the parallel coefficients for the control groups. It is essential that practice on a criterion task differentially influence the abilities of the individuals in the sample. Without these differential influences, we can expect changes in means but not stabilities as a result of practice on a related task. The explanation requires that we also observe changes in rank orders of individuals on the ability dimensions. Without these changes in rank orders, the observed decrements in predictive validity coefficients would not be observed. Differential influences on stability/reliability coefficients were not observed.

A second explanation for the observed predictive validity decrements was based on regression to the mean hypotheses. This explanation assumes that individuals hired into organizations are normally selected from the upper parts of relevant ability dimensions. Once these individuals are selected and trained, they exhibit the expected regression to the ability means of the population from which they were selected. This regression to the mean of the high ability population should result in significant changes in rank order along the ability dimension and decreasing relations between initial ability measures and performance later in practice or training with those highest on the ability dimensions used for selection exhibiting the greatest amount of regression. This regression to the mean explanation is important because

it suggests that the decrements in predictive validities that have been observed do not result in decrements in the utility of the ability tests used as selection devices. That is, the mean performance on jobs or tasks for highly selected groups would be higher than the mean performance on the same jobs or tasks for unselected groups, thus preserving the utility of the selection devices. But, the relationship between initial ability measures and performance, within the selected group, would decline.

A test of this hypothesis would require a highly selected and a randomly selected group be given extensive training on a criterion task or a job. The highly selected group should show a regression to a higher ability mean than the unselected group as a function of time and/or practice on the task. This was done in the AIC study reported above. There were no observable differential regression effects for the high ability and randomly selected ability groups. The explanation for predictive validity decrements based on regression to the mean of highly selected population for the high ability selected groups was not supported in these data.

Hypotheses derived from neither the regression to the mean nor the transfer of training explanation for predictive validity decrements were supported. However, the data did offer support for several general propositions that might be derived from implications of the generally observed superdiagonal matrix of relations among measures of performance or ability measures. Implications of the observed decrements in predictive validity were also explored. The support offered was found in the general increase in variance accounted for in criterion task performance when measures of performance assessed during training were entered into multiple regression equations. The results were consistent and conclusive. Training measures entered in the equations after either pre- or post-training ability measures, accounted for significant and practically meaningful amounts of additional variance in the criterion task. When measures of training performance were entered first, the pretest or posttest ability measures did not account for significant amounts of additional variance. These improvements in variance accounted for were general across the high and random ability groups. Although the relations involving the random ability groups were frequently stronger than those based on the high ability groups because of the enhanced variance in these groups, the results were general across both samples. The clear and consistent message from these results is that predictions of skill acquisition and performance were improved by the use of the information contained in training performance assessments. The use of these training assessments to improve the validity of prediction equations in an operational environment should have an important effect on the utility of any selection program in which performance on complex criterion tasks must be predicted. The gains in utility are especially important in operational environments where the costs of errors or poor judgments during later stages of training when the individuals may be performing the task with only minimal supervision or during operational stages when the new employees are completely on their own may be extreme. The gains in predictive validity should more than offset the costs of quantifying and assessing training performance. The use of early performance measures to update predictions of later performance is operationally feasible. When training is divided into different and distinct phases, admission into later stages of training can be made dependent on performance levels in initial, and likely less

expensive, stages. Sequential selection procedures need not depend solely on ability measures administered in predetermined orders; they can also make use of valid information available in the measures of training performance.

The results shown in Figure 1 were not predicted by the authors. They are, however, potentially important for sorting out the complexities that prevent simple explanations and easily implemented practical solutions to the problems raised by decreasing predictive validity coefficients. The regression of the predictive validities of written pretest ability measures for blocks of trials in the AIC task had a slope of $-.77$ ($p < .01$). The negative slope of the written pretests validities onto the blocks of trials replicates the documented decrement in predictive validities that was the impetus for these studies. However, the regression of the predictive validities of the computerized pretests onto the blocks of trials revealed no trend. The meta-analysis reported by Hulin et al. (1990) was based mainly on the results of traditional, non-computerized, ability tests. Thus, this portion of the study supports their conclusions. However, the flat slope of the regression of the predictive validities of the computerized pretests raises several issues that cannot be addressed in this report. It does suggest that our attempts to explain the predictive validity decrements should perhaps shift somewhat from a search for a general explanation to one that focusses on examinations of what classifications of tests reveal validity decrements or increments under what conditions and with what criterion tasks. Additional attempts to explain this observed difference in the slopes of predictive validities onto blocks of training trials would represent little more than speculation. Such explanation must await additional data gathered specifically to examine differences between different classifications of ability tests.

The regression of postdictive validities of the written posttests for the same blocks of trials was not significantly different from the slope for the written pretests. The regression of the postdictive validities of the computerized posttests onto the blocks of trials in the AIC task was not significantly different from the slope for the computerized pretests. The similarity of the slopes of these two groups of pretests and posttests confirm the earlier hierarchical regression analyses of pretests, posttests, and training measures entered in various combinations and orders as predictors of criterion task performance.

The apparent generality of the phenomenon of predictive validity decrements described by Hulin et al. (1990) may be less than originally suggested. The past literature on the topic was based almost exclusively on written ability tests, job samples, apparatus tests of psychomotor skills, and measures of performance on related tasks (e.g., undergraduate grades as predictors of performance in graduate and professional schools). Computerized tests that take advantage of the flexibility of technology, that contain elements of speededness and dynamic presentations of stimulus materials, may offer assessment techniques that change significantly the factorial composition of the resulting measures. The addition of unmeasured specific or group factors that exploit communalities among items and measures that depend on factors not present in more traditional ability assessments, may change in subtle but important ways the factorial composition of the resulting ability measures.

This hypothesis remains to be explored; it does offer avenues for research not systematically considered in the literature.

A logical and necessary next step is a thorough and systematic facet analysis of the likely components of variance contained in traditional and computerized tests, and how these components of variance might be expected to influence predictive validities across time. Such an analysis is beyond the scope of this report.

REFERENCES

- Adams, J.A. (1957). The relationship between certain measures of ability and the acquisition of a psychomotor response. Journal of General Psychology, 56, 121-134.
- Alvares, K.M., & Hulin, C.L. (1972). Two explanations of temporal changes in ability-skill relationships: A literature review and a theoretical analysis. Human Factors, 14, 295-308.
- Alvares, K.M., & Hulin, C.L. (1973). Changes in ability measures as a function of complex skill acquisition. Organizational Behavior and Human Performance, 9, 169-185.
- Dunham, R.B. (1974). Ability-skill relationships: An empirical explanation of change over time. Organizational Behavior and Human Performance, 12, 372-382.
- French, J.W. (1954). Manual for kit of selection tests for reference aptitude and achievement factors. Princeton, N.J.: Educational Testing Service.
- Hulin, C.L., Henry, R.A., & Noon, S.L. (1990). Adding a Dimension: Time as a factor in the generalizability of predictive relationships. Psychological Bulletin, 107, 328-340.
- Humphreys, L.G. (1960). Investigations of the simplex. Psychometrika, 4, 313-323.
- Humphreys, L.G. (1973). Postdiction of the GRE and eight semesters of college grades. Journal of Educational Measurement, 10, 179-184.
- Humphreys, L.G., & Lin, P. (1977). Prediction of academic performance in graduate and professional school. Applied Psychological Measurement, 1, 249-257.
- Humphreys, L.G., & Taber, T. (1973). Postdiction study of the GRE and eight semesters of college grades. Journal of Educational Measurement, 10, 179-184.
- Ramey, C.T. & Ramey, S.L. (1990). Intensive educational interventions for children of poverty, Intelligence, 14, 1-9.

- Schneider, W., Vidulich, M., & Yeh, Y. (1982). Training spatial skills for air-traffic control. Proceedings of the Human Factors Society, 10-14.
- Vidulich, M., Yeh, Y., & Schneider, W. (1983). Time-compressed components for air-intercept control skills. Proceedings of the Human Factors Society, 161-164.
- Zigler, E. (1987). Early experience, malleability, and Head Start. In J.J. Gallagher & C.T. Ramey (Eds), The Malleability of Children, pp. 85-95. Baltimore, MD: Paul H. Brookes.